# 大規模ブログデータベースにおける単語の出現頻度の解析

## —ノイズ成分の数理構造解析から食の流行把握への応用まで—

渡邊隼史　社会データ構造化センター（情報システム研究機構）　特任助教
h.wata@ism.ac.jp
**arXiv:1604.00762** (2016)

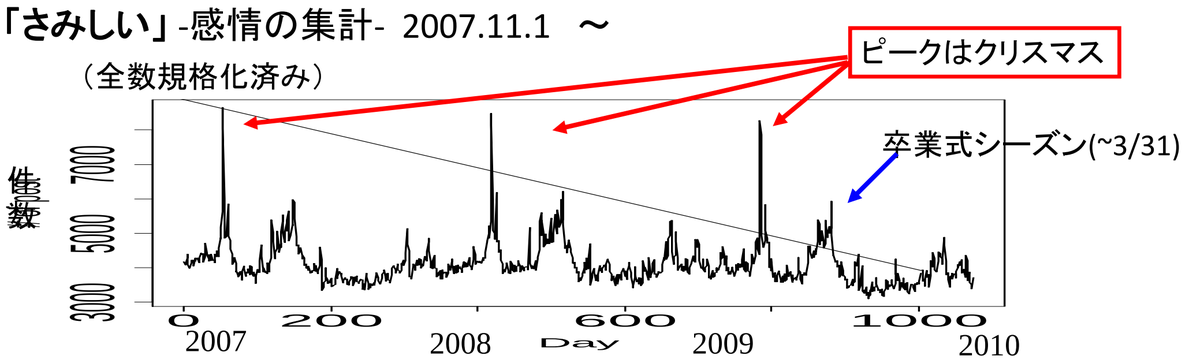
### 【ブログデータ】

- Web上の日記のようなもの。 日付付きのテキストデータ
- 2006年11月1日～現在までの日本語ブログ記事30億記事を利用
- データ収集は、ホットリンク社のロコミ係長を利用
- 基本的な量である着目キーワードの国内ブログ上での出現頻度時系列に着目

国内主要ブログパイダを網羅

### 「さみしい」-感情の集計- 2007.11.1 ～

（全数規格化済み）




→集団としての人間の活動や感情を量的にとらえることができる。

### 【ブログデータの応用研究： 食の流行把握】

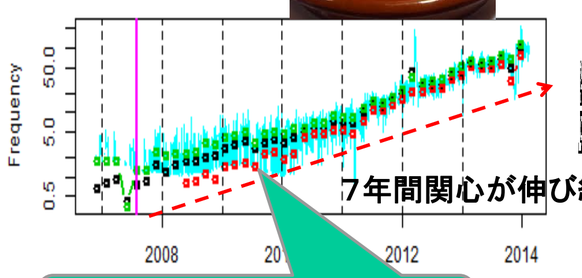
共同研究：小森、榊(ホットリンク)

食の流行の現状把握・予兆発見

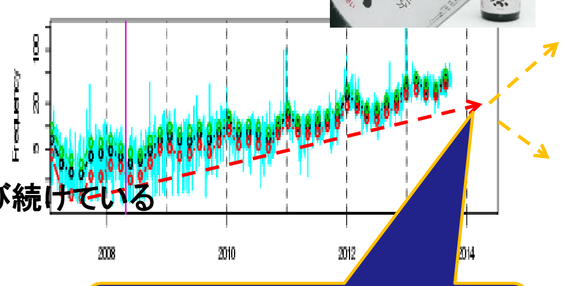
それをソーシャルメディアでできないか？

**ニーズ (by 大手食品メーカー, by 雑誌流行)**
消費者の動向の変化を他より早く知りたい
雑誌やTVにないような流行の兆しを知りたい

- アヒージョ




- 獺祭(だっさい)




7年間関心が伸び続けている

関心の増加を
ここできづけたらいいな…

今のブームの過熱って
どのくらい？ 未来は？

食の事項から、人知れず、関心が増えてるものを探す
→食の辞書を作成し，件数が多くなく，長期間上昇し続けてる語を抽出

「食」語を教える
（1）単語の区切り
（2）食べ物かどうか？

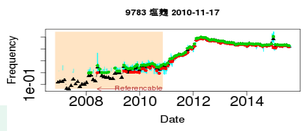
- 主に、以下の2つの技術を開発：

1. 食べ物に関する新語候補発見技術： 新しい食べ物語が次々登場への対応
   **「ブーム」と書かれたブログ記事に含まれる単語から新着の食べ物語を抽出**
   複合語処理（専門用語抽出）->「食ワード」と「非食ワード」の分類（LSA）
   (1)味噌カレー牛乳ラーメン、オーガニックエクストラバージンココナッツオイル，昼飲み
   (2)裏難波、キャロラインリーパー（唐辛子），ファスティング（断食）
2. 時系列継続的な関心増加（減少）の検出 **：多様な時系列にロバストに統一的な処理の必要**
   長期的な単語の書き込み件数の上昇・下降の把握
   瞬間的増加率でなく伸びの継続，**急増語**では予兆としては遅い．

2010年11月の関心連続増加ランキング(件数 20件以下)

その後関心が1年以上伸び続けた語

| | | | | |
|---|---|---|---|---|
| オムツケーキ | 48 | 3.25 | 16.4 | 82 |
| ケークサレ | 41 | 4.97 | 19.3 | 47 |
| パイセン | 48 | 2.20 | 7.37 | 96 |
| 川越達也 | 39 | 7.70 | 14.2 | 47 |
| 赤飯さん | 31 | 3.49 | 19.6 | 34 |
| 旦那さん弁当 | 49 | 2.10 | 6.96 | 77 |
| 変形フレンチ | 47 | 2.02 | 11.9 | 78 |
| かりんとうまんじゅう | 38 | 3.93 | 5.94 | 42 |
| SABON | 35 | 1.71 | 19.3 | 53 |
| 塚田農場 | 44 | 3.32 | 4.43 | 82 |
| かりんとう饅頭 | 30 | 4.23 | 8.01 | 43 |
| はま寿司 | 47 | 2.57 | 6.79 | 94 |
| 糖質制限 | 49 | 1.36 | 11.1 | 104 |
| アジング | 49 | 1.28 | 17.5 | 86 |
| グリーンスムージー | 38 | 5.18 | 12.8 | 61 |
| 着丼 | 33 | 6.35 | 4.54 | 88 |
| 炭酸パック | 33 | 2.98 | 17.61 | 44 |



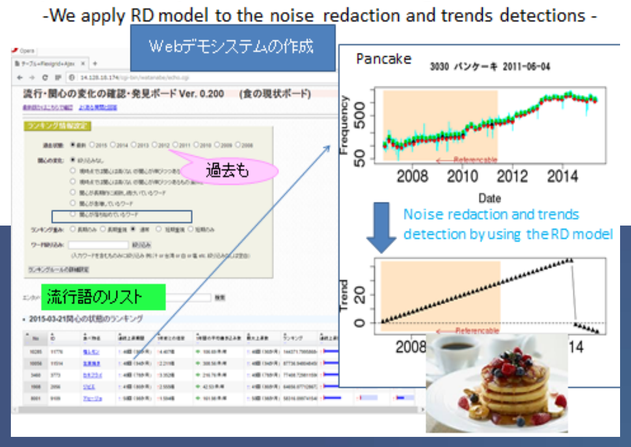

| | | | | |
|---|---|---|---|---|
| 朝ラー | | | | |
| 濃厚つけ麺 | | | | |
| 森崎友紀 | | | | |
| 川越シェフ | | | | |
| 焼き小籠包 | 29 | 4.74 | 7.25 | 45 |
| 塩麹 | 36 | 1.95 | 4.91 | 56 |
| アヒージョ | 48 | 1.79 | 6.06 | 105 |
| バーニャカウダソース | 45 | 1.66 | 7.80 | 49 |
| マカロンタワー | 49 | 1.61 | 5.82 | 106 |
| アイシングクッキー | 41 | 1.89 | 11.9 | 98 |
| 酵素ドリンク | 28 | 4.37 | 9.16 | 68 |
| 料理男子 | 29 | 2.87 | 2.06 | 51 |
| ホワイトフレンチ | 49 | 1.70 | 4.87 | 86 |
| メガポテト | 47 | 1.67 | 3.54 | 48 |
| バラ焼き | 45 | 1.43 | 5.54 | 49 |
| A5ランク | | | | |
| 餡掛け | | | | |
| トンテキ | | | | |
| アガベシロップ | | | | |
| ダイパーケーキ | | | | |
| サムギョプル | | | | |
| おんどる | | | | |
| 塩唐揚げ | | | | |
| お家ごはん | 43 | 1.31 | 12.7 | 68 |
| とんてき | 49 | 1.25 | 5.56 | 65 |
| まぜそば | | | | |
| 牛骨ラーメン | | | | |
| クリームチーク | | | | |
| 玉ねぎスープ | | | | |
| プレモル | | | | |
| チーズトッポギ | | | | |
| 米粉100% | | | | |


列：食べ物名，連続上昇月数，年増加率，年平均件数
(2010.11時点)，歴代最大連続上昇月数(2015.11時点)

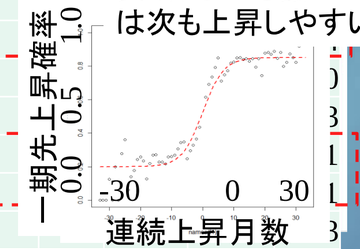
**Application2: Systematic detection of vogues of foods**

-We apply RD model to the noise redaction and trends detections -

Webデモシステムの作成





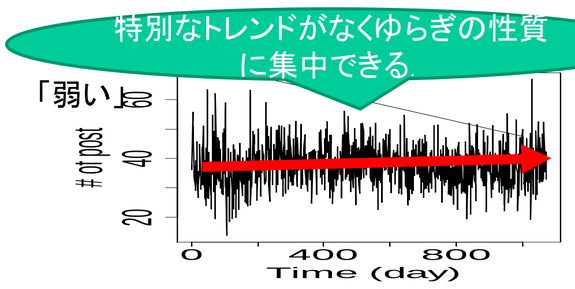



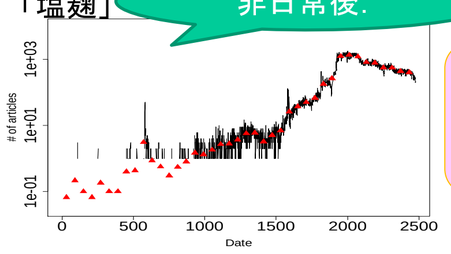

長く伸びてる語
は次も上昇しやすい？

### 【ブログデータの基礎研究：なにもイベントない語のゆらぎの性質】

共同研究：佐野(筑波)、高安（ソニーCSL）、高安（東工大）

特別なトレンドがなくゆらぎの性質に集中できる




特別をよく知るため、まず通常を知る

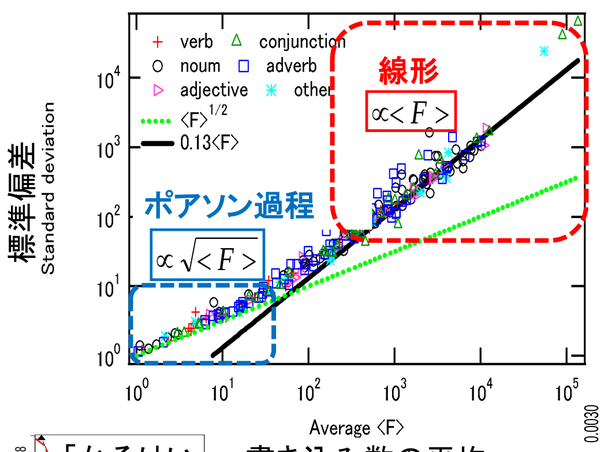
—ブログにおけるゆらぎの（時間）スケーリングとは：

- 平均書き込み数とゆらぎ(標準偏差)の間のべき乗関係 （ゆらぎのスケーリング FS）

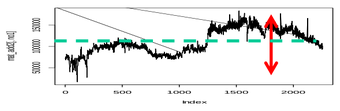
（1）平均が小さい領域では0．5乗（ポアソン的） $\sigma_F \propto \sqrt{\mu_F}$

**（2）平均が大きいときは比例 （非ポアソン的）** $\sigma_F \propto \mu_F$

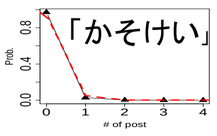
あるキーワードが1日当たり日本中のブログに何回かかれるか？



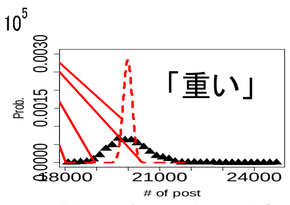

様々な言葉の書き込み数の
**平均書き込み数** $\mu_F$
**とゆらぎの大きさ** $\sigma_F$
**の関係を調べた**



非ポアソン的




書き込み数の平均




黒点：データ
赤線：ポアソン分布

珍しい単語 → 頻出する単語（流量の大きい川に対応）

### 【ブログデータのゆらぎの性質： 拡張ランダム拡散モデル】

$$P(F_i(t);c_i) = \int \Phi(\varepsilon - x(t);t) \exp(-c_i \varepsilon) \frac{[c_i \varepsilon]^F}{F_i(t)!} d\varepsilon$$

非定常拡張

$$P(F_i(t);c_i) = \int \Phi(\varepsilon - x(t);t) \exp(-c_i(t) \varepsilon) \frac{[c_i(t) \varepsilon]^F}{F_i(t)!} d\varepsilon$$

ゆらぎのスケーリングを説明するモデル [S.Meloni *et al* PRL2008, Y. Sano et.al 2012 JEIC]
川の流量、インターネットパケットetc、複雑ネットワーク上での輸送ゆらぎの平均場近似

- **ポアソン分布の重ねあわせ**
  - 個々のブロガーがランダムにポアソン分布で書き込む効果
  - と社会全体のブロガー数がゆらぐ効果の、
  - 重さねあわせとして、書き込み数の分布決定

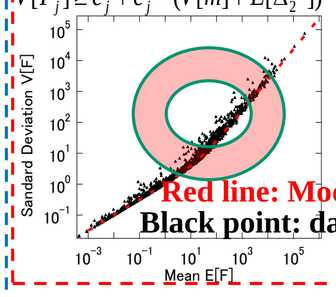
**Fi(t):単語i書き込み数**
**cj(t):単語iの書き込み数平均**
**x(t):規格化システムサイズ**
**ε(t): システムサイズの乱数成分**

**拡張ランダム拡散モデル(非定常)**

ランダム拡散モデル(定常)

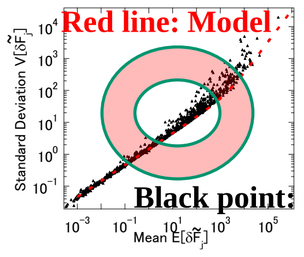
Temporal FS

$V[F_j] \geq c_j + c_j^2 \cdot (V[m] + E[\Delta_2^2])$



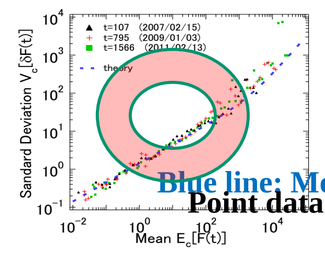

Scaled temporal FS

$V[\delta\tilde{F}_j] \geq 2 \cdot E[\frac{1}{m}] \cdot c_j + c_j^2 (E[\Delta_2^2] \cdot E[\frac{1}{m^2}])$



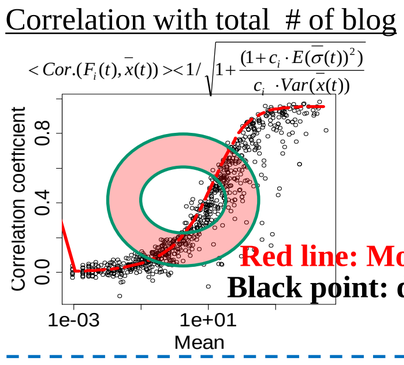

Ensemble FS

$V_c[\delta\tilde{F}(t)] \geq 2 \cdot (E_c[F(t)]) + 2 \cdot (E_c[F(t)]^2) \cdot (\Delta_0^2)$



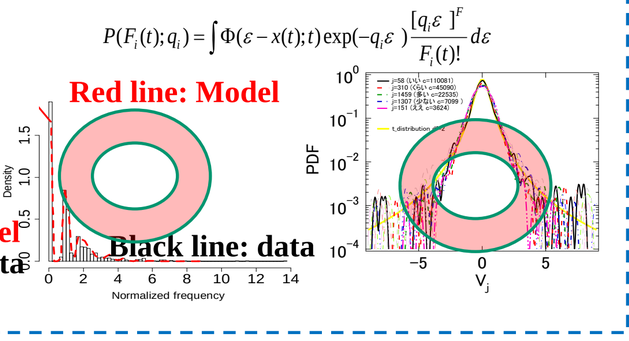

Correlation with total # of blog

$< Cor.(F_i(t), \bar{x}(t)) > < 1 / \sqrt{1 + \frac{(1 + c_i \cdot E(\bar{\sigma}(t))^2)}{c_i \cdot Var(\bar{x}(t))}}$



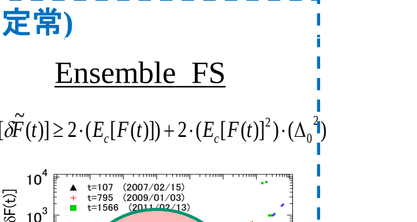

Probability density function

$P(F_i(t);q_i) = \int \Phi(\varepsilon - x(t);t) \exp(-q_i \varepsilon) \frac{[q_i \varepsilon]^F}{F_i(t)!} d\varepsilon$

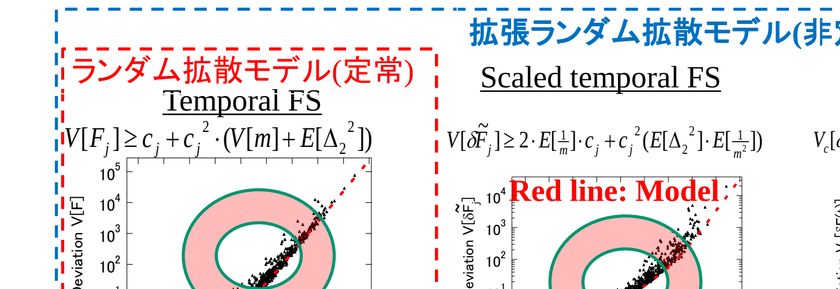



### 【ブログデータのゆらぎの性質と拡張ランダム拡散モデルの応用】

差分の確率密度関数の再現




10万語/日から0.016語/日 まで、
11桁のスケールを一つ変数
（時間平均）の変化のみで再現










**形容詞使用の異常度でみる国民的イベント**